)les on the validity of epidemiologic
idies. Am J Epidemiol 1984;120:

Savitz DA, Jöckel K-H, et al. Effects
ential exposure misclassification in
dies. Am J Epidemiol 1992;135:

, Brenner H. Correcting for nondif-
classification in ecologic analyses.
93;42:117–26.

Bergsten JW, Alexander BV, et al.
urveys of indoor ^{222}Rn. Health Phys
-6.

chwehr MB, Nazaroff WW, et al.
of airborne radon-222 concen-
US homes. Science 1986;234:

earch Council, Committee on the
fects of Ionizing Radiation. Health
n and other internally deposited
s. Washington, DC: National Acad-
EIR IV, 1988.

iggins CL, Humble CG, et al. Ciga-
and lung cancer in New Mexico.
ir Dis 1988;137:1110–3.

Inc. SAS/IML software: usage and
rsion 6, First Edition. Cary, NC:
Inc, 1989.

don levels in United States homes
counties. Health Phys 1991;60:

lt of Health and Human Services.
health consequences of smoking,
rogress: a report of the Surgeon
ville, MD: USGPO, 1989. (DHHS
. (CDC) 89-8411).

n and lung cancer. J Natl Cancer
45–57.

cil on Radiation Protection and
Evaluation of occupational and
exposure to radon and radon
ie United States. Bethesda, MD:
NCRP report no. 78.)

ommission on Radiological Pro-
ancer risk from indoor exposures
lters. Oxford: Pergamon Press,
blication no. 50.)

American Journal of Epidemiology

Vol. 139, No. 3
Printed in U.S.A.

Energy Adjustment Methods for Nutritional Epidemiology: The Effect of Categorization

Charles C. Brown,[1] Victor Kipnis,[1] Laurence S. Freedman,[1] Anne M. Hartman,[2] Arthur Schatzkin,[3] and Sholom Wacholder[4]

The authors discuss the interpretation of four alternative energy adjustment methods (Residual, Standard, Partition, and Nutrient Density) that have been proposed for the analysis of nutritional epidemiology studies. These methods have so far been compared under circumstances where intake of the nutrient of interest is measured as a continuous variable. Because it is common practice to categorize nutrient intakes in the analysis, the authors investigate the effect of such categorization on the interpretation of results from the four methods with the use of computer simulations and statistical theory. They consider four cases: where the nutrient intake is either divided into quartiles or ordered so as to investigate trend over the quartile groups, combined with using an adjusting variable that is either continuous or categorized. The results show: 1) the Residual, Standard, and Partition methods are no longer equivalent as they are in the continuous case; 2) compared with the Standard method, the Residual method appears to be more powerful for detecting trends in relative odds, is more robust to residual confounding when the adjustment variable is categorized, and provides more meaningful odds ratios; and 3) the Residual and Nutrient Density methods give closely similar results. *Am J Epidemiol* 1994;139:323–38.

confounding factors (epidemiology); diet; epidemiologic methods; models, statistical; nutrition assessment

There have recently appeared several articles and letters concerning the statistical methods to be used for appropriately analyzing relations between certain nutrient intakes and a specific disease (1–5). Several statistical methods have been proposed and their relations discussed. The methods that we will consider in this article are:

the Standard method (4),

$$\text{logit}(P) = \beta_{0S} + \beta_{1S}F + \beta_{2S}T,$$

the Residual method (1),

$$\text{logit}(P) = \beta_{0R} + \beta_{1R}R + \beta_{2R}T,$$

the Energy Partition method (3),

$$\text{logit}(P) = \beta_{0P} + \beta_{1P}F + \beta_{2P}(T - F),$$

and the Nutrient Density method (4),

$$\text{logit}(P) = \beta_{0N} + \beta_{1N}(F/T) + \beta_{2N}T,$$

where P is the probability of disease, T is the total energy intake, and F is the intake of the nutrient of interest, which for the purposes of this paper we will assume to be fat. The intakes T and F are measured in kilocalories per day. The variable R in the Residual

Received for publication December 15, 1992, and in final form October 11, 1993.

[1] Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.
[2] Applied Research Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.
[3] Cancer Prevention Studies Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.
[4] Biostatistics Branch, Division of Cancer Etiology, National Cancer Institute, Bethesda, MD.
Reprint requests to Dr. Charles C. Brown, Biometry Branch, DCPC, National Cancer Institute, Executive Plaza North, Suite 344, Bethesda, MD 20892.

method is the "energy-adjusted fat intake," that is, the residual from regressing F on T (1).

In a previous article (5), we have discussed the equivalences between the first three of these methods and have pointed out that the estimates of the parameters in any one of these three methods can be used to determine directly the estimates of the parameters in the other two. Understanding these equivalences relieves some of the concern over the choice of which method to use and redirects attention to the question regarding the specific effects one wishes to estimate.

In practice, epidemiologists usually analyze continuous data, such as nutrient intakes, by categorizing subjects into a small number (3–5) of groups according to their intake and establishing relative odds for the different groups. Thus, although the methods listed above have formed the basis of the debate, they are not the methods that are used in practice. Unfortunately, the equivalences between the first three methods no longer exist once the data have been categorized. In this paper, we explore the consequences that categorization has for analyses based on these methods. We start from the premise that the true model linking disease and dietary intake is linear in the continuous variables, as indicated by the models above, and we use computer simulations and statistical theory to study the impact of categorization on the results of analyses.

MATERIALS AND METHODS

We used computer simulation techniques to compare the four logistic regression methods for the analysis of diet-disease data. We simulated the results of a series of random case-control studies having N_1 cases and N_2 controls generated from a linear logistic regression model in which the logit of the probability of being a case is linearly dependent on fat intake and on intake from non-fat sources. For each study subject ($i = 1, \ldots, N_1 + N_2$), we randomly generated a pair of dietary intake variables, F_i (fat kilocalories) and T_i (total kilocalories) from a bivariate normal distribution with mean fat kilocalories = 930, mean total kilocalories = 2,380, fat kilocalories standard deviation = 290, total kilocalories standard deviation = 550, and correlation between fat and total kilocalories = 0.85. These values were derived from data collected in a dietary methodology study that was conducted as part of the US-Finland Lung Cancer Prevention Study (6). Letting β and γ represent the disease effects of a fat kilocalorie and a non-fat kilocalorie, respectively, the disease-score $S_i = \beta F_i + \gamma (T_i - F_i)$ was compared with a randomly generated Z_i (uniformly distributed between 0 and 1) to classify the subjects into cases and controls. The N_1 subjects with the largest $S_i - \text{logit}(Z_i)$ were classified as cases, and the remaining N_2 subjects were classified as controls.

Additional regression covariates were computed as the non-fat variable $T_i - F_i$, the fat density variable $D_i = F_i/T_i$ and the fat residual variable $R_i = F_i - \hat{F}_i = F_i - (\hat{\alpha} + \hat{\beta}T_i)$, where $\hat{\alpha}$ and $\hat{\beta}$ denote the least squares estimates from a linear regression of F on T for all $N_1 + N_2$ subjects. For each dietary variable ($x = F, T, T - F, D$ and R), quartile indicators were formed in the following manner by using all $N_1 + N_2$ subjects:

$$I_{1x} = 1 \text{ if } x \text{ falls into the first quartile; 0 otherwise}$$
$$I_{2x} = 1 \text{ if } x \text{ falls into the second quartile; 0 otherwise}$$
$$I_{3x} = 1 \text{ if } x \text{ falls into the third quartile; 0 otherwise}$$
$$I_{4x} = 1 \text{ if } x \text{ falls into the fourth quartile; 0 otherwise}$$

A trend variable was computed as $I_{Tx} = I_{1x} + 2I_{2x} + 3I_{3x} + 4I_{4x}$.

Then, the following series of four logistic regression models were evaluated for each of the four analysis methods on each simulated case-control study,

$$\text{logit}[P] = \alpha + \beta I_{Tx} + \gamma y \qquad \text{Trend-Continuous}$$

$$\alpha + \beta I_{Tx} + \sum_{i=2}^{4} \gamma_i I_{iy} \qquad \text{Trend-Quartile}$$

$$\alpha + \sum_{i=2}^{4} \beta_i I_{ix} + \gamma y \qquad \text{Quartile-Continuous}$$

$$\alpha + \sum_{i=2}^{4} \beta_i I_{ix} + \sum_{i=2}^{4} \gamma_i I_{iy} \qquad \text{Quartile-Quartile}$$

where x and y denote the nutrient intake variable of interest and adjustment variable, respectively.

For the Residual method, $x = R$ and $y = T$; for the Standard method, $x = F$ and $y = T$; for the Partition method, $x = F$ and $y = T - F$; and for the Density method, $x = D$ and $y = T$. In the Quartile-Continuous and Quartile-Quartile models β_i represents the log relative odds of disease in the ith ($i = 2,3,4$) quartile relative to the first quartile. In the Trend-Continuous and Trend-Quartile models, β represents the trend in log relative odds over the four quartiles.

To further understand the effect of categorization, we developed theoretic results for simple linear regression models using the Residual, Standard, and Partition parameterizations. The theory deals with the case of linear approximations to stochastic regression models, and is outlined in the Appendix. Results from the theory are based on asymptotic arguments and therefore apply to large samples.

RESULTS

Table 1 presents the results of 1,000 simulations of fitting the Quartile-Continuous regression models to data from a case-control study of 100 cases and 100 controls for four sets of coefficients (per 1,000 kilocalories): 1) only intake from fat is related to disease, fat coefficient = 2.4 and non-fat coefficient = 0; 2) intake from fat imparts an extra risk above that from non-fat, fat coefficient = 1.8 and non-fat coefficient = 0.6; 3) the source of energy is not related to risk, fat coefficient = non-fat coefficient = 1.2; and 4) intake from fat increases risk while intake from non-fat is protective, fat coefficient = 1.8 and non-fat coefficient = −0.6. The table contains the average estimated log odds ratio, the average standard error of the log odds ratio and the proportion of estimated log odds ratios statistically significant at the 5 percent level. The empirical standard deviation of the 1,000 estimated log odds ratios is not included in the table because it agrees closely with the average standard errors. The results of these simulations indicate the following:

- The Residual and Density methods give nearly identical results for all four sets of coefficients.
- Except when energy source is unrelated to risk, the Standard method estimates higher log odds ratios than does the Residual method: for example, in case 1 the average Residual method estimates are 0.333, 0.571, and 0.919, while the average Standard method estimates are 0.495, 0.836, and 1.303.
- Except when intake from non-fat is protective, the Partition method estimates higher log odds ratios than does the Standard method; in case 1, the average Partition method

TABLE 1. Average log odds ratio ± average log odds standard error (proportion significant at $p \leq 0.05$); Quartile-Continuous model (from 1,000 simulations)

Method	Quartile 2 vs. Quartile 1	Quartile 3 vs. Quartile 1	Quartile 4 vs. Quartile 1
	Fat coefficient = 2.4, non-fat coefficient = 0.0		
Residual	0.333 ± 0.427 (12.5%)	0.571 ± 0.427 (25.8%)	0.919 ± 0.432 (57.2%)
Standard	0.495 ± 0.470 (19.9%)	0.836 ± 0.541 (35.4%)	1.303 ± 0.688 (47.6%)
Partition	0.629 ± 0.436 (31.4%)	1.064 ± 0.451 (66.2%)	1.667 ± 0.499 (93.2%)
Density	0.341 ± 0.436 (13.8%)	0.573 ± 0.439 (26.2%)	0.924 ± 0.438 (54.5%)
	Fat coefficient = 1.8, non-fat coefficient = 0.6		
Residual	0.162 ± 0.425 (6.5%)	0.281 ± 0.425 (9.3%)	0.452 ± 0.426 (19.7%)
Standard	0.240 ± 0.467 (8.9%)	0.410 ± 0.536 (12.2%)	0.634 ± 0.677 (17.1%)
Partition	0.467 ± 0.433 (19.7%)	0.799 ± 0.445 (43.5%)	1.252 ± 0.486 (72.5%)
Density	0.175 ± 0.433 (6.8%)	0.282 ± 0.437 (9.1%)	0.456 ± 0.433 (19.7%)
	Fat coefficient = 1.2, non-fat coefficient = 1.2		
Residual	−0.009 ± 0.426 (5.1%)	−0.008 ± 0.426 (5.7%)	−0.016 ± 0.425 (5.5%)
Standard	−0.009 ± 0.467 (5.3%)	−0.012 ± 0.536 (4.5%)	−0.040 ± 0.675 (6.2%)
Partition	0.310 ± 0.433 (12.3%)	0.537 ± 0.444 (22.9%)	0.831 ± 0.481 (42.5%)
Density	0.008 ± 0.434 (5.7%)	−0.012 ± 0.439 (5.1%)	−0.012 ± 0.434 (5.5%)
	Fat coefficient = 1.8, non-fat coefficient = −0.6		
Residual	0.336 ± 0.414 (11.9%)	0.598 ± 0.414 (30.3%)	0.920 ± 0.419 (61.0%)
Standard	0.488 ± 0.457 (18.0%)	0.822 ± 0.528 (33.6%)	1.301 ± 0.671 (49.9%)
Partition	0.473 ± 0.423 (19.5%)	0.796 ± 0.439 (43.5%)	1.261 ± 0.481 (75.3%)
Density	0.331 ± 0.423 (11.1%)	0.604 ± 0.427 (28.7%)	0.924 ± 0.426 (60.1%)

estimates are 0.629, 1.064, and 1.667 compared to the average Standard method estimates of 0.495, 0.836, and 1.303; however, in case 4, when non-fat is protective, the average Partition method estimates are 0.473, 0.796, and 1.261, while the average Standard method estimates are 0.488, 0.822, and 1.301.

- For the Residual and Density methods, the standard errors of the estimated log odds ratios are constant across quartiles, while they are increasing for the Standard and Partition methods (much more strongly for the Standard method).
- In case 3, the Residual, Standard, and Density methods each find statistically significant log odds ratios in approximately 5 percent of the simulations.
- The Partition method more frequently finds statistically significant log odds ratios than do the other three methods: for example, comparing quartile 4 with quartile 1 in case 1, the Partition method finds 93.2 percent of the simulations statistically significant while the Residual, Standard, and Density methods find only 57.2 percent, 47.6 percent, and 54.5 percent significant.
- The Residual method finds more statistically significant log odds ratios than the Standard method when comparing the extreme quartiles 4 versus 1, while the reverse is seen when comparing quartile 2 with quartile 1 and comparing quartile 3 with quartile 1.

Additional information is displayed in figure 1, which presents scatterplots of the estimated log odds ratios comparing quartile 4 with quartile 1 for 100 simulations of case 1. Estimates are in very close agreement between the Standard and Partition methods (correlation coefficient, $r = 0.91$) and the Residual and Density methods ($r = 0.90$), while estimates correlate less closely between the Residual and Standard methods ($r = 0.71$) and the Residual and Partition methods ($r = 0.71$). A consistent pattern of correlations is observed for all four cases and for each log odds ratio. The Standard-Partition and Residual-Density correlations were generally much greater than the Residual-Standard and Residual-Partition correlations.

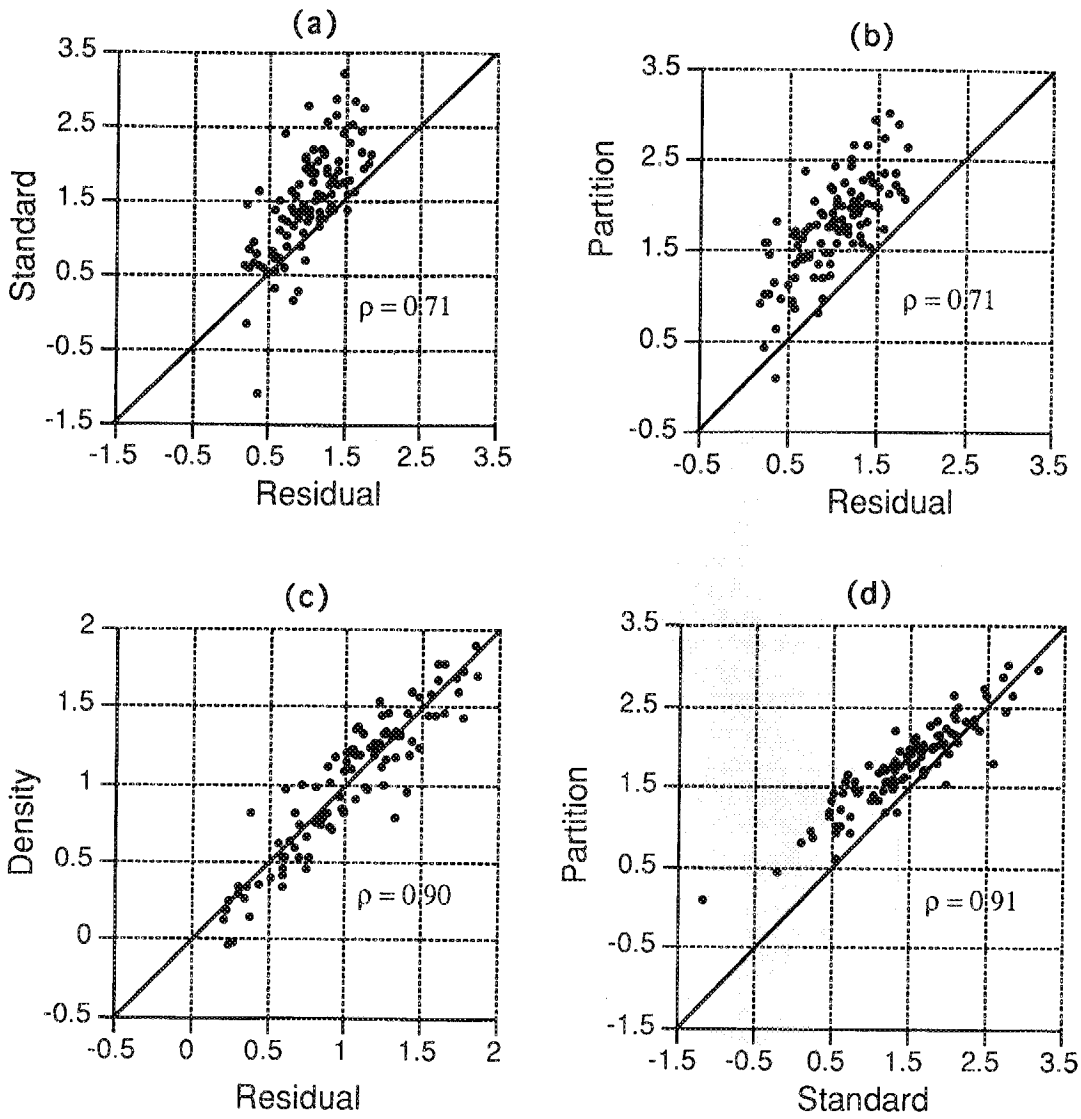


FIGURE 1. Comparison of estimated log relative odds: quartile 4 vs. quartile 1: Quartile-Continuous model; 100 simulations of case 1, fat coefficient = 2.4, non-fat coefficient = 0.

Similar to table 1, the results of fitting the Quartile-Quartile regression models are presented in table 2. The conclusions are similar to those for the Quartile-Continuous models except that:

- The average log odds ratio estimates from the Standard, Partition, and Density methods are greater in table 2 than in table 1; for example, comparing quartile 4 with quartile 1 in case 1, the average Standard method log odds ratio estimate from the Quartile-Continuous model is 1.303 and is 1.523 from the Quartile-Quartile model.
- The Standard, Partition, and Density methods have more statistically significant estimated log odds ratios in table 2 than in table 1; for example, comparing quartile 4 with quartile 1 in case 1, the Standard method finds 47.6 percent statistically significant using the Quartile-Continuous model and 61.5 percent statistically significant using the Quartile-Quartile model.

TABLE 2. Average log odds ratio ± average log odds standard error (proportion significant at $p \leq 0.05$); Quartile-Quartile model (from 1,000 simulations)

Method	Quartile 2 vs. Quartile 1	Quartile 3 vs. Quartile 1	Quartile 4 vs. Quartile 1
	Fat coefficient = 2.4, non-fat coefficient = 0.0		
Residual	0.332 ± 0.430 (12.7%)	0.572 ± 0.430 (25.9%)	0.920 ± 0.435 (57.1%)
Standard	0.582 ± 0.492 (23.1%)	0.977 ± 0.570 (40.8%)	1.523 ± 0.680 (61.5%)
Partition	0.643 ± 0.441 (32.6%)	1.090 ± 0.455 (67.4%)	1.710 ± 0.498 (94.4%)
Density	0.381 ± 0.439 (15.3%)	0.627 ± 0.442 (29.7%)	0.963 ± 0.441 (58.5%)
	Fat coefficient = 1.8, non-fat coefficient = 0.6		
Residual	0.161 ± 0.428 (6.2%)	0.282 ± 0.428 (9.0%)	0.452 ± 0.428 (19.8%)
Standard	0.353 ± 0.488 (11.9%)	0.560 ± 0.563 (17.1%)	0.892 ± 0.669 (30.0%)
Partition	0.494 ± 0.437 (21.0%)	0.842 ± 0.449 (47.3%)	1.323 ± 0.486 (77.9%)
Density	0.217 ± 0.436 (8.2%)	0.339 ± 0.439 (11.7%)	0.497 ± 0.435 (22.4%)
	Fat coefficient = 1.2, non-fat coefficient = 1.2		
Residual	−0.009 ± 0.428 (5.6%)	−0.007 ± 0.428 (6.0%)	−0.015 ± 0.427 (5.5%)
Standard	0.131 ± 0.489 (7.2%)	0.155 ± 0.564 (6.1%)	0.259 ± 0.669 (7.6%)
Partition	0.348 ± 0.437 (14.4%)	0.594 ± 0.448 (26.4%)	0.927 ± 0.480 (51.2%)
Density	0.055 ± 0.436 (5.7%)	0.050 ± 0.440 (5.7%)	0.033 ± 0.435 (5.4%)
	Fat coefficient = 1.8, non-fat coefficient = −0.6		
Residual	0.338 ± 0.418 (12.1%)	0.603 ± 0.418 (29.5%)	0.928 ± 0.423 (59.7%)
Standard	0.514 ± 0.481 (18.5%)	0.903 ± 0.563 (36.5%)	1.403 ± 0.669 (54.7%)
Partition	0.467 ± 0.427 (19.1%)	0.790 ± 0.443 (42.0%)	1.248 ± 0.479 (76.3%)
Density	0.351 ± 0.427 (12.5%)	0.633 ± 0.431 (31.8%)	0.949 ± 0.429 (61.3%)

Table 3 and figure 2 present the results of fitting the trend regression models, Trend-Continuous and Trend-Quartile. The table contains the average estimated slope, its average estimated standard error, and the proportion of estimated trends statistically significant at the 5 percent level. As for the estimated log odds ratios, the variance of the

TABLE 3. Average fat trend slope ± average slope standard error (proportion significant at $p \leq 0.05$), from 1,000 simulations

Method	Trend-Continuous model	Trend-Quartile model
	Fat coefficient = 2.4, non-fat coefficient = 0.0	
Residual	0.304 ± 0.135 (61.4%)	0.305 ± 0.136 (61.1%)
Standard	0.426 ± 0.217 (49.7%)	0.496 ± 0.217 (62.6%)
Partition	0.537 ± 0.156 (94.7%)	0.548 ± 0.156 (95.4%)
Density	0.306 ± 0.137 (60.1%)	0.318 ± 0.138 (64.6%)
	Fat coefficient = 1.8, non-fat coefficient = 0.6	
Residual	0.155 ± 0.134 (20.3%)	0.154 ± 0.135 (19.3%)
Standard	0.215 ± 0.214 (16.5%)	0.295 ± 0.213 (30.6%)
Partition	0.406 ± 0.152 (77.1%)	0.426 ± 0.152 (81.9%)
Density	0.156 ± 0.136 (20.6%)	0.169 ± 0.136 (23.5%)
	Fat coefficient = 1.2, non-fat coefficient = 1.2	
Residual	−0.001 ± 0.134 (5.9%)	−0.001 ± 0.134 (6.1%)
Standard	−0.005 ± 0.214 (5.1%)	0.085 ± 0.213 (7.6%)
Partition	0.267 ± 0.151 (42.3%)	0.295 ± 0.151 (50.6%)
Density	−0.001 ± 0.136 (5.1%)	0.013 ± 0.137 (5.5%)
	Fat coefficient = 1.8, non-fat coefficient = −0.6	
Residual	0.300 ± 0.131 (62.7%)	0.302 ± 0.133 (62.6%)
Standard	0.415 ± 0.213 (49.1%)	0.452 ± 0.214 (57.7%)
Partition	0.406 ± 0.152 (77.4%)	0.402 ± 0.151 (76.8%)
Density	0.302 ± 0.133 (62.6%)	0.310 ± 0.135 (64.6%)

estimated trend slopes is not included because the average trend slope standard error is only slightly smaller (1 percent) than the standard deviation of the 1,000 estimated slopes. The results of the simulations indicate the following conclusions:

- The Residual and Density methods give essentially the same results in all four cases.
- On average, the ordering of estimated slopes are Partition > Standard > Residual; for example, the case 1 Trend-Continuous model results are 0.537 > 0.426 > 0.304.
- Using the Trend-Continuous model, the Residual method finds more statistically significant estimated trend slopes than does the Standard method, 61.4 percent versus 49.7 percent in case 1.
- In general, using the Standard, Partition, and Density methods, the estimated trend slope and proportion of statistically significant estimates from the Trend-Quartile model are both greater than those from the Trend-Continuous model.

We found good agreement between the theory developed in the Appendix and the results of these simulations. For example, table 4 provides the theoretical expected values and standard errors of the log odds ratios for the Trend-Continuous models to compare with those computed from the simulations.

DISCUSSION

Previous work on the Residual, Standard, and Partition methods with the nutrient (e.g., fat) intake expressed as a continuous variable has revealed that the meaning of the fat coefficient varies according to the method (5). For both the Residual and Standard methods, the fat coefficient represents the effect on disease of increasing fat intake by *substituting* fat for non-fat nutrients; for the Partition method, it represents the effect of *adding* fat to the diet. Thus, one would expect estimated coefficients for fat to differ according to the method used, except that estimated coefficients from the Residual and Standard methods should agree. However, as observed by Kushi et al. (7) and seen in our simulations, different estimated relative odds are obtained from the Residual and Standard methods in which fat intake is discretized.

We will first discuss the case where the variable used for adjustment (e.g., total energy intake in the Standard method) is kept as a continuous variable (table 1 and table 3, column 1).

TABLE 4. Asymptotic mean fat trend slope ± asymptotic standard error: Trend-Continuous model

Method	Trend-Continuous model
	Fat coefficient = 2.4, non-fat coefficient = 0.0
Residual	0.303 ± 0.127
Standard	0.419 ± 0.206
Partition	0.547 ± 0.146
	Fat coefficient = 1.8, non-fat coefficient = 0.6
Residual	0.152 ± 0.127
Standard	0.210 ± 0.206
Partition	0.410 ± 0.146
	Fat coefficient = 1.2, non-fat coefficient = 1.2
Residual	0.000 ± 0.127
Standard	0.000 ± 0.206
Partition	0.273 ± 0.146
	Fat coefficient = 1.8, non-fat coefficient = −0.6
Residual	0.303 ± 0.127
Standard	0.419 ± 0.206
Partition	0.410 ± 0.146

Residual versus Standard method

The reason for the discrepancy between the Residual and Standard method results can be understood by considering the meaning of the relative odds parameters resulting from categorization of the fat variable in these two methods. Consider the risk of a subject in the second quartile relative to a subject in the first quartile. The Standard method's relative odds represent the effect on disease of substituting enough fat for non-fat in a subject's diet so that the subject will be taken out of the first into the second

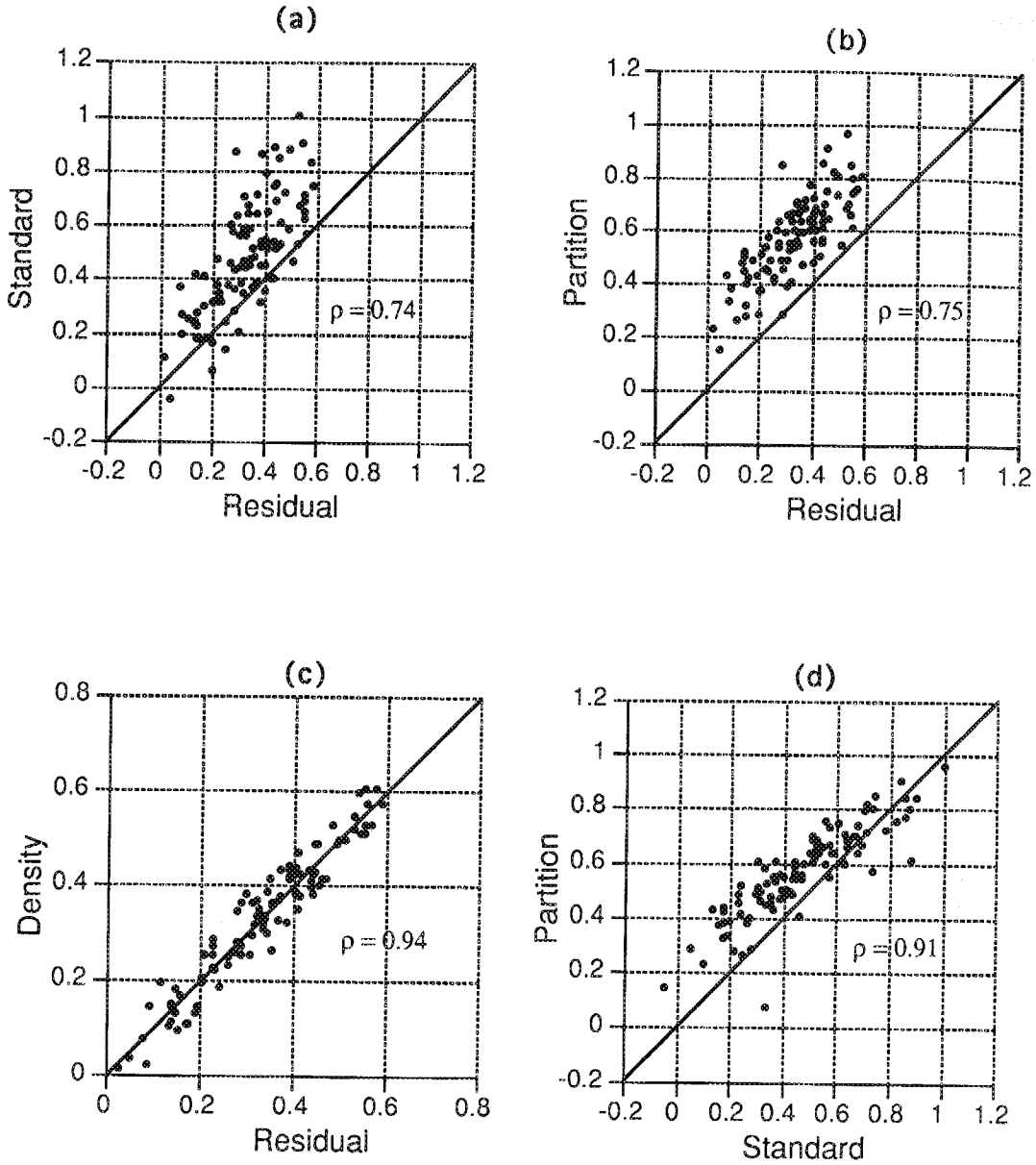
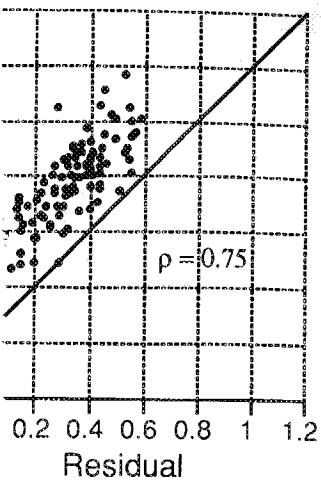


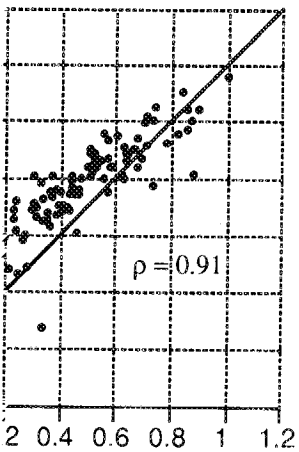
FIGURE 2. Comparison of estimated log relative odds trend slopes: Trend-Continuous model; 100 simulations of case 1, fat coefficient = 2.4, non-fat coefficient = 0.

quartile of *absolute fat intake*. On the other hand, the Residual method's relative odds represent the effect of substituting enough fat for non-fat to move the subject from the first to the second quartile of *residual fat intake*. Because the between-subject variation in fat residual is smaller than the between-subject variation in absolute fat intake (4), a smaller increase in fat intake is required to induce a step-up in fat residual quartile than is required to induce a step-up in absolute fat quartile. In other words, the difference in fat intake between subjects in the first and second quartiles of absolute fat is greater than that between subjects in the corresponding quartiles of fat residual. For our simulations, the average intake of kilocalories from fat was 565, 835, 1,023, and 1,299 for subjects in each quartile of absolute fat intake and 739, 879, 981, and 1,121

(b)



(d)



inuous model; 100 simulations of

quired to induce a step-up
iartile. In other words, the
intake between subjects in
nd quartiles of absolute fat
at between subjects in the
iartiles of fat residual. For
the average intake of kilo-
was 565, 835, 1,023, and
s in each quartile of abso-
1 739, 879, 981, and 1,121

for subjects in each quartile of residual fat intake. Therefore, when fat intake is categorized, the relative odds are expected to be larger for the Standard method than for the Residual method. This difference emphasizes that the relative odds from the two methods no longer estimate the same quantity.

Even though the relative odds from the Standard method tend to be greater than those from the Residual method, the observed ordering may be reversed, as noted by Kushi et al. (7) and seen in figure 1a. This lack of concordance between the two sets of estimates can be explained by noting that categorizing subjects by absolute fat intake may give quite different results than categorization by residual fat intake. For a correlation between intake from fat and total energy intake of 0.85, table 5 shows how the proportions of subjects within each quartile of absolute fat intake are expected to be distributed among the quartiles of residual fat intake. For example, of those falling into the second quartile of absolute fat intake, 28 percent, 30 percent, 26 percent, and 16 percent are expected to fall into the first through fourth quartiles of fat residual. Figure 3 illustrates this phenomenon for 100 randomly generated sets of values for fat and total intake (see also Kushi et al. (7)). As a result, for any single set of data, the Standard method and Residual method relative odds estimates may be based on two substantially different sets of individuals.

Our simulations and theory also allowed us to examine the statistical power of these methods for detecting a relative odds trend. The larger relative odds obtained using the Standard method do not translate into a power advantage over the Residual method. When adjustment for total energy intake is accomplished by including total kilocalories as a continuous covariate, the Residual method has greater power than the Standard method. This is because categorization affects not only the magnitude of the relative odds trend but also the magnitude of its standard error. The Standard method's standard error is larger than the Residual method's standard error and by a factor greater than the ratio of the relative odds trends of the two methods (see Appendix). Figure 3 shows why this occurs. Because of the high correlation between total energy and fat intake, nearly all subjects having the lowest total energy intake will fall into the quartile of lowest fat intake and nearly all subjects having the highest total energy intake will fall into the quartile of highest fat intake. Therefore, these subjects will contribute very little information regarding the fat relative odds, particularly that of quartile 4 versus quartile 1. This loss of information translates into the larger standard error for the Standard method relative odds estimates as shown in tables 1–3. Because the power for detecting a relative odds trend is a function of the ratio of its expected value to its standard error, the Residual method ends up giving a higher power than the Standard method.

TABLE 5. Percent of subjects in absolute fat intake quartiles expected to fall into residual fat intake quartiles*

Absolute fat quartile	Residual fat quartile				Total
	1	2	3	4	
1	50	28	16	6	100
2	28	30	26	16	100
3	16	26	30	28	100
4	6	16	28	50	100

* Correlation (fat calories, total calories) = 0.85.

Standard versus Partition method

Comparison of the results from the Standard method with those from the Partition method are also of interest. Both the Standard and Partition methods use absolute level as the fat intake variable and thus a change from the first to the second quartile represents the same magnitude of change. However, as noted earlier, the fat coefficient in the Partition method represents the effect of adding dietary fat, whereas the fat coefficient in the Standard method represents the effect of substituting fat for non-fat. Because these effects are not usually the same, one should not expect to obtain the same estimates from these methods. The result of substituting fat for non-fat depends on the

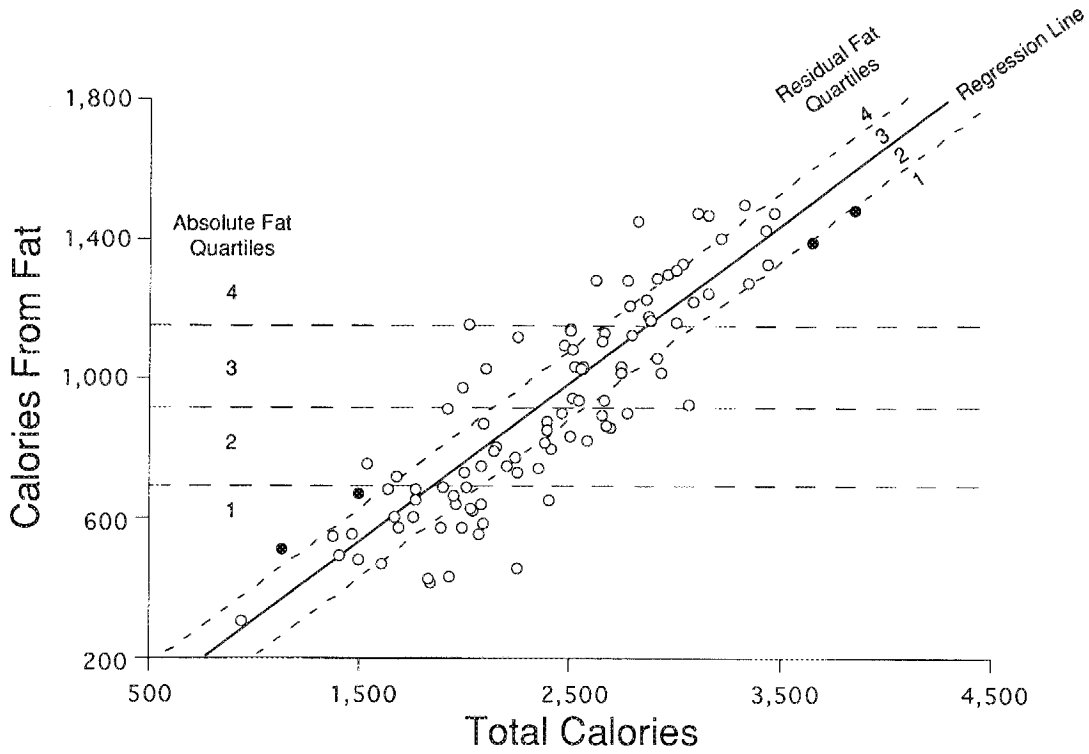


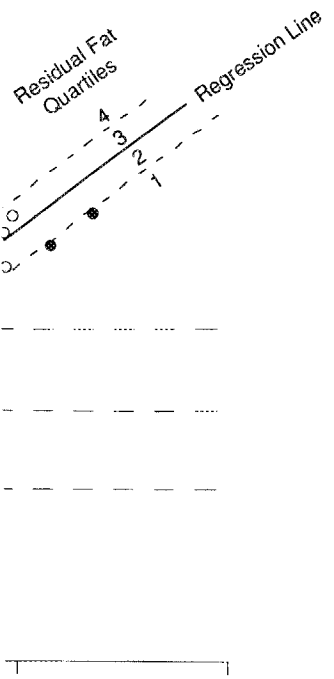
FIGURE 3. Example of 100 random dietary intakes categorized by quartiles of absolute fat intake (horizontal dashed lines) and quartiles of residual fat intake (angled dashed lines). Correlation of fat calories with total calories = 0.85. Dark points fall into opposite extreme quartiles.

effect of non-fat intake as well as on the effect of fat consumption. In our simulations, the Partition method gave larger relative odds than the Standard method when non-fat intake was assumed to increase the risk of disease. For the simulations where non-fat intake had a negative (i.e., protective) effect against disease, the Partition method relative odds tended to be smaller than those from the Standard method. In the case where non-fat had zero effect on disease, the Partition method again gave estimated relative odds larger than the Standard method, although using continuous variables the relative odds would have been equal. Categorization of the fat variable results in attenuation of the relative odds when the true relation with disease risk is linear with fat intake on the continuous scale. The degree of attenuation, however, depends on the correlation of fat intake with the adjustment variable in the model. When the correlation is high, as in the Standard method, then the attenuation is greater than when the correlation is lower, as in the Partition method (see Appendix). Hence, the Partition method estimated relative odds tend to be larger.

It is also notable that the Standard and Partition methods give results that are closely correlated (figure 1d). As mentioned above, relative odds from the Standard method may be larger or smaller than those from the Partition method, depending on the size and direction of the effect of non-fat energy on disease. However, for a fixed effect of non-fat intake, the scatterplot shown in figure 1d indicates that one should be able to closely predict the Standard method relative odds for fat from the relative odds using the Partition method.

Nutrient Density method

The Nutrient Density method fat coefficient represents the effect of increasing the percentage of fat in the diet while keeping



of absolute fat intake (horizontal
n of fat calories with total calories

on is greater than when the
wer, as in the Partition
ndix). Hence, the Partition
l relative odds tend to be

ble that the Standard and
ls give results that are
(figure 1d). As mentioned
odds from the Standard
rger or smaller than those
method, depending on the
n of the effect of non-fat
. However, for a fixed ef-
ake, the scatterplot shown
tes that one should be able
the Standard method rela-
om the relative odds using
iod.

method

)ensity method fat coeffi-
ie effect of increasing the
in the diet while keeping

total energy intake constant. The Nutrient Density method therefore represents another version of a substitution method. The Residual and Nutrient Density methods appear to produce results that both agree well on average (tables 1 and 3) and also agree consistently across different data sets (figures 1c and 2c). The empirical results indicate that these methods are closely linked, at least in the simulation cases that we considered. Kushi et al. (7) also report close agreement between the results from these two methods.

Categorizing the adjusting variable

The discussion so far has covered the case where we have modeled the adjusting variable (non-fat intake for the Partition method and total energy intake for the other three methods) as a continuous variable assumed to have a linear effect. The results in table 2 and table 3, column 2, cover analyses where we modeled the adjusting variable as categorical (in quartiles). These results demonstrate that when the true relation between disease and the adjusting variable is linear, then adjustment using categorization may produce quite different results from those using the correct regression model. This bias is due to "residual confounding" by the adjusting variable (8). When the adjusting variable is categorized, its confounding effect is not completely captured by the regression model. The magnitude of the remaining bias depends on the strength of correlation of the fat variable with the adjusting variable and the strength of the adjusting variable-disease relation. Therefore, the ordering of the fat-adjusting variable correlations implies that the residual confounding bias should be largest for the Standard method and smallest for the Residual method. Both the Nutrient Density and the Partition methods suffer from this bias; however, its magnitude is considerably smaller than for the Standard method because the fat-adjusting variable correlation in these methods is less strong.

When the adjusting variable is quartiles of total energy intake, the biases produce a spurious statistical power advantage for the Standard method over the Residual method. The same bias leads to statistically significant results in greater than 5 percent of the simulations for case 3, effect of fat = effect of non-fat = 1.2, where there is really no effect of *substituting* fat calories for non-fat calories as implicitly defined by the Standard and Residual methods. Proper adjustment by a model containing a linear effect of total energy intake results in, as expected, approximately 5 percent of the simulations statistically significant for the Residual and Standard methods (table 1), whereas when adjustment is by quartiles of kilocalories, the bias increases the percentage of significant results using the Standard method to nearly 7 percent (table 2).

Conclusion

The results presented in this paper are restricted to categorization of variables into quartiles. Other simulations and our theory indicate that the same qualitative effects occur when smaller or larger numbers of categories are used. However, as the number of categories increases, the results approach those based on the continuous variable models when the assumed model is true.

This study comprises results from simulations of a logistic linear model relating dietary intake to disease incidence and from theory based on a multiple linear regression model. The simulations are limited to a rather simple model and only four sets of parameter values. We developed the theory partly to gain insight into some of the surprising aspects of the simulation results and partly to understand which patterns seen in the simulations would apply more generally and which were specific to the simulation cases chosen. That we were able to obtain formulas from a multiple linear regression model that could predict rather well the results of logistic regression simulations gives us confidence that the patterns we have described are quite general. However, we should emphasize that our models are simpler than those encountered in real epidemiologic studies, for example, containing

no risk factors that are confounded with dietary intake, no measurement error, a linear relation between fat intake and total energy intake, and a linear relation between disease and dietary intake. Nevertheless, we believe the study has led to some useful insights. The most important message is that categorizing continuous variables in a statistical analysis is not without consequences. The Standard and Residual method differences in relative odds estimates and statistical powers are due entirely to categorization. The relative odds estimates and powers are identical using the same methods with continuous variables.

Which of the two methods, Residual or Standard, is preferable, given that one wishes to categorize? In this work, we have found that the Residual method apparently provides more statistical power for detecting trends in relative odds when total energy intake is correctly specified in the model. In addition, the Residual method is more robust to residual confounding when the total energy variable is not specified correctly in the model. We also believe the relative odds estimated from the Residual method better reflect the effect of nutrient substitution than those from the Standard method. The Residual method quartiles are based on the variability of residual fat intake, which is equivalent to the variability in absolute fat intake for subjects *with the same total energy intake*. On the other hand, the Standard method quartiles are based on the variability of absolute fat intake for the entire population *regardless of their total energy intake*. Both methods are concerned with the effect of substituting fat for non-fat, keeping total energy intake constant. Use of quartiles of residual fat intake is more consistent with estimating this substitution effect. It therefore appears to us that the Residual method carries several advantages over the Standard method when categorization is used.

It has been suggested that the Standard method Quartile-Quartile model should not be used when the correlation between the nutrient and total energy intakes is 0.8 or greater. This is because within the lowest quartile of total energy there would be few or no subjects in the highest quartile of absolute fat and within the highest quartile of energy few would fall into the lowest quartile of absolute fat. Therefore, it is suggested that a comparison of the highest and lowest quartiles of absolute fat would be too imprecise if the analysis were stratified by quartiles of total energy. However, rather than stratification for energy adjustment, the Standard method Quartile-Quartile regression model produces increased precision (over stratification estimates) by using information from all the other interquartile comparisons. For example, table 2 shows that the Standard method Quartile 4 versus Quartile 1 comparison has, on average, only a 38 percent larger standard error than the Quartile 2 versus Quartile 1 comparison.

Another message of the paper is that the Residual and the Nutrient Density methods appear to give very similar results. The reasons for this near equivalence need to be better understood.

Because the Partition coefficients have a fundamentally different interpretation from the Residual or Standard coefficients, we are not surprised by the differences found in this study in its relative odds and statistical powers. This study reinforces the message that these two methods, by answering different questions, may lead to quite different numerical results (5). One needs to decide whether one is interested in the effect of adding fat intake to the diet or in the effect of increasing fat consumption by substituting fat for non-fat intake, and *then* choose the appropriate method.

Categorization is the epidemiologists' protection against gross misspecification of the model. It is indeed a useful device, but as shown in this paper (and in a recent publication on nondifferential misclassification (9)), its consequences on the results of an analysis can be surprising. Investigators should not assume that the statistical properties of a model with continuous variables will necessarily transfer across to the same model with the variables categorized.

REFERENCES

1. Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 1986;124:17–27.
2. Pike MC, Bernstein L, Peters RK. Re: "Total energy intake: implications for epidemiologic analyses." (Letter). Am J Epidemiol 1989;129: 1312–13.
3. Howe GR. The first author replies. (Letter). Am J Epidemiol 1989;129:1314–15.
4. Willett W. Nutritional epidemiology. New York: Oxford University Press, 1990.
5. Kipnis V, Freedman LS, Brown CC, et al. Interpretation of energy adjustment models for nutritional epidemiology. Am J Epidemiol 1993;137: 1376–80.
6. Pietinen P, Hartman AM, Haapa E, et al. Reproducibility and validity of dietary assessment instruments. I. A self-administered food use questionnaire with a portion size picture booklet. Am J Epidemiol 1988;128:655–66.
7. Kushi LH, Sellers TA, Potter JD, et al. Dietary fat and postmenopausal breast cancer. J Natl Cancer Inst 1992;84:1092–9.
8. Savitz DA, Barón AE. Estimating and correcting for confounder misclassification. Am J Epidemiol 1989;129:1062–71.
9. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 1991;134:1233–44.

APPENDIX

Statistical theory for regression coefficients of categorized variables

Let y be a continuous variable representing "level" of disease (note that in the main text the symbol y is used differently). Consider the linear stochastic regression model linking disease to continuous nutrient intakes $\mathbf{x} = (x_1, \ldots, x_k)$

$$y = \mathrm{E}(y \mid \mathbf{x}) + \epsilon = \mathbf{x}\beta + \epsilon, \tag{A1}$$

where $\beta = (\beta_1, \ldots, \beta_k)'$ is the $(k \times 1)$ vector of regression coefficients corresponding to each nutrient intake, and ϵ is a disturbance term such that $\mathrm{E}(\epsilon \mid \mathbf{x}) = 0$ and $\mathrm{E}(\epsilon^2 \mid \mathbf{x}) = \sigma^2$. By centralizing y and the nutrient intakes $\mathbf{x}$, we can assume that $\mathrm{E}(y)$ and $\mathrm{E}(\mathbf{x}) = \mathbf{0}$. Let $\mathbf{z} = (z_1, \ldots, z_m)$ be a $(1 \times m)$ vector of transformed nutrient variables,

$$z_j = g_j(\mathbf{x}), j = 1, \ldots, m,$$

where transformations g_j can be, for example, categorization into quartile indicator variables, or to a variable representing a trend over categories, as described earlier. We will assume that the transformed variables are centralized so that

$$\mathrm{E}(z_j) = 0, j = 1, \ldots, m.$$

The regression $\mathrm{E}(y \mid \mathbf{z})$ may not be linear because of the transformation, so we will consider the mean-square linear regression of y on $\mathbf{z}$, that is the least squares linear approximation of y by the linear combination $\mathbf{z}\gamma$, where $\gamma = (\gamma_1, \ldots, \gamma_m)'$. In the framework of the main text the values $\gamma_1, \ldots, \gamma_m$ will correspond to regression coefficients in models linking disease to categorical nutrient intakes. We have

$$y = \mathbf{z}\gamma + \delta,$$

where

$$\mathrm{E}(\delta) = 0 \text{ and } \mathrm{E}(\mathbf{z}'\delta) = 0.$$

With n subjects in the sample, let $\hat{\gamma}$ be the ordinary least squares estimator of γ, the vector of regression coefficients. We can show that for large n

$$\sqrt{n}\,(\hat{\gamma} - \gamma) \text{ is approximately distributed as } N(\mathbf{0}, \mathbf{V}^{-1}\mathrm{E}(\mathbf{z}'\delta^2\mathbf{z})\mathbf{V}^{-1}), \tag{A2}$$

where $\mathbf{V} = \mathrm{E}(\mathbf{z}'\mathbf{z})$ is the variance-covariance matrix of vector $\mathbf{z}'$. Then, by definition, the asymptotic mean and variance-covariance matrix of γ are as follows:

$$\mathrm{E}(\hat{\gamma}) = \gamma = \mathbf{V}^{-1}E(\mathbf{z}'\mathbf{x})\beta \tag{A3}$$